In [1]:

```
import pandas as pd
import requests
from bs4 import BeautifulSoup
import requests
from selenium import webdriver
import re
import numpy as np
from sklearn.utils import shuffle
from time import sleep
```

```
/home/helmi/.local/lib/python3.5/site-packages/requests/__init__.py:91: RequestsDependencyWarni
ng: urllib3 (1.24.1) or chardet (3.0.4) doesn't match a supported version!
  RequestsDependencyWarning)
```

In [2]:

```
from selenium.webdriver.common.keys import Keys
```

Click the AFFICHER TOUS LES ARTICLES button to display all products in every category

In [3]:

```
#For each categorie we need to show all the product by invoking the "AFFICHER TOUS LES ARTICLES" button.
def afficher_tous(current_driver):
    button=current_driver.find_element_by_link_text("AFFICHER TOUS LES ARTICLES")
    button.click()
    sleep(7)#5
    elm=current_driver.find_element_by_tag_name("html")
    elm.send_keys(Keys.END)
    sleep(5)#20
    elm.send_keys(Keys.HOME)
    # store it to string variable
    page_source = current_driver.page_source
    current_soup=BeautifulSoup(page_source,'html5lib')
    return current_soup
```

Get the driver with selenium

In [4]:

```
def get_driver(url):#"https://www.evaps.fr/boutique.html"
    driver = webdriver.Chrome('./chromedriver')
    sleep(4) #10
    driver.get(url)
    sleep(4)#10
    return driver
```

In [5]:

```
driver_cigarette=get_driver("https://www.evaps.fr/boutique.html")
```

In [6]:

```
soup=afficher_tous(driver_cigarette)
```

Fetch the detail page for every categorie to get the: -Price, -Description and -Image

In [7]:

```
def detail_page(soup):    # soup for which categorie  (detail page for every given categorie)
    all_devs=links=soup.select('div[class="infos-box"]')
    all_links=[l.select_one('a') for l in all_devs]
    hrefs=["https://www.evaps.fr/"+l.get("href") for l in all_links]
    return hrefs
```

Fetch all the 3 categories:

*E-cigarette

*E-liquide

*DIY

In [8]:

```
def go_categorie_xpath(xpath):#"//*[@id='menu2']/a"
    href=driver_cigarette.find_element_by_xpath(xpath).get_attribute("href")
    cat_driver=get_driver(href)
    cat_soup=afficher_tous(cat_driver)
    #cat_driver.close()*****************************
    return cat_soup
```

Names and Brands

In [9]:

```
def fill_features(which_soup):
    all_lab=which_soup.select('div[class="infos-box"]')
    labs=[n.select('a[class="libelle"]') for n in all_lab ]
    all_names=[n[0].get_text() for n in labs]
    list_brand=[b.split("-",1)[1] if b.find("-")!=-1 else np.nan for b in all_names]
    return {"Names":all_names, "Brands":list_brand}
```

Fetch data from Detail page

In [12]:

```
def list_deatil(soup):
    hrefs=detail_page(soup)
    liste_prix=list()
    liste_imgs=list()
    liste_desc=list()
    def list_prix():
        for h in hrefs:
            print("href : "+h)
            p=requests.get(h)
            soup_alterna=BeautifulSoup(p.text,'html5lib')
            prix=soup_alterna.select('span[itemprop="price"]')[0].get_text()
            liste_prix.append(prix)
            a=h.split("details-produit.",1)[1]
            b="https://www.evaps.fr/documents/media/images/contenu/"+a
            c=b.replace("html","jpg")
            liste_imgs.append(c)
            try:
                desc=soup_alterna.select_one("div[id='mini-description'] p").get_text()
            except:
                desc=np.nan
            liste_desc.append(desc)

    list_prix()
    return {"list_prix":liste_prix,"Photo":liste_imgs,"Desc":liste_desc}
```

All the work is done here

In [13]:

```
def work():
    list_brand=list()
    soup_diy=go_categorie_xpath('//*[@id="menu12"]/a')
    soup_elquide=go_categorie_xpath("//*[@id='menu2']/a")

    all_prix_ecigarette=list_deatil(soup)['list_prix']
    #print("Liste prix "+all_prix_ecigarette[0])

    all_prix_eliquide=list_deatil(soup_elquide)['list_prix']
    all_prix_diy=list_deatil(soup_diy)['list_prix']

    #******************Photo************************

    all_photo_ecigarette=list_deatil(soup)["Photo"]
    all_photo_eliquide=list_deatil(soup_elquide)["Photo"]
    all_photo_diy=list_deatil(soup_diy)["Photo"]

    #**************DEscription*********************************
    all_d_ecigarette=list_deatil(soup)["Desc"]
    all_d_eliquide=list_deatil(soup_elquide)["Desc"]
    all_d_diy=list_deatil(soup_diy)["Desc"]
    #************************NAMES*************************
    names_e_cigarette=fill_features(soup)["Names"]
    names_e_liquide=fill_features(soup_elquide)["Names"]
    names_diy=fill_features(soup_diy)["Names"]

    #******************BRAND*********************************
    marque_e_cigarette=fill_features(soup)["Brands"]
    marque_e_liquide=fill_features(soup_elquide)["Brands"]
    marque_e_diy=fill_features(soup_diy)["Brands"]


    #***************Categorie*********************
    categorie_cigarette=["e_cigarette"]*len(names_e_cigarette)
    categorie_liquide=["e_liquide"]*len(names_e_liquide)
    categorie_diy=["diy"]*len(names_diy)


    Names=names_e_cigarette+names_e_liquide+names_diy
    Brands=marque_e_cigarette+marque_e_liquide+marque_e_diy
    Prices=all_prix_ecigarette+all_prix_eliquide+all_prix_diy
    Categories=categorie_cigarette+categorie_liquide+categorie_diy
    Photos=all_photo_ecigarette+all_photo_eliquide+all_photo_diy
    Descrs=all_d_ecigarette+all_d_eliquide+all_d_diy

    df=pd.DataFrame(
    {
        "Name":Names,
        "Price":Prices,
        "Brand":Brands,
        "Categorie":Categories,
        "Photo":Photos,
        "Description":Descrs

    },
)
    return  shuffle(df)
    driver_cigarette.close()
```

In [ ]:

```
df=work()
```

In [15]:

```
df["Photo"].isnull().all()
```

Out[15]:

```
False
```

Convert dataframe to csv

In [ ]:

```
df.to_csv("csv_data.csv")
```

Convert dataframe to json

In [21]:

```
df.to_json("json_data.json")
```

In [22]:

```
df.head()
```

Out[22]:

	Brand	Categorie	Description	Name	Photo	Price
102	CoilArt	e_cigarette	NaN	Kit Mage Mech Tricker - CoilArt	https://www.evaps.fr/documents/media/images/co...	69.90
77	Priv 230W - Smoktech	e_cigarette	Laissez place à la somptueuse Box S Priv de Sm...	Box S-Priv 230W - Smoktech	https://www.evaps.fr/documents/media/images/co...	56.90
82	Vaporesso	e_cigarette	Succombez au redoutable charme du kit Switcher...	Kit Switcher 220W - Vaporesso	https://www.evaps.fr/documents/media/images/co...	79.90
250	Petit Nuage	e_liquide	Le eliquide Flocon Pressé Petit Nuage 60ml ❤ e...	Flocon Pressé 60ml - Petit Nuage	https://www.evaps.fr/documents/media/images/co...	24.90
109	Eleaf	e_cigarette	NaN	Ikonn Total / Ello Mini XL - Eleaf	https://www.evaps.fr/documents/media/images/co...	43.90

Convert ipynb to Python script and to pdf

In [26]:

```
! ipython nbconvert --to html evaps.ipynb
```

```
[TerminalIPythonApp] WARNING | Subcommand `ipython nbconvert` is deprecated and will be removed
 in future versions.
[TerminalIPythonApp] WARNING | You likely want to use `jupyter nbconvert` in the future
/home/helmi/.local/lib/python3.5/site-packages/requests/__init__.py:91: RequestsDependencyWarni
ng: urllib3 (1.24.1) or chardet (3.0.4) doesn't match a supported version!
  RequestsDependencyWarning)
[NbConvertApp] Converting notebook evaps.ipynb to html
[NbConvertApp] Writing 311322 bytes to evaps.html
```

In [29]:

```
! wkhtmltopdf evaps.html evaps.pdf
```

```
Loading page (1/2)
Warning: Failed to load file:///home/helmi/Desktop/Web_scraping/custom.css (ignore)
Printing pages (2/2)
Done
```